# Webarchivering

Voerman, Gerrit; Voorburg, René; Huurdeman, Hugo

*Published in:*
Archievenblad



*Publication date:*
2012

# Webarchivering: een pleidooi voor een archiverend netwerk van organisaties

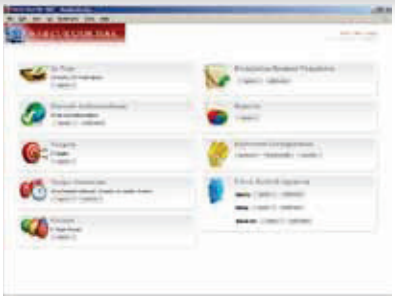
Gerrit Voerman, René Voorburg en Hugo Huurdeman ■

**Het world wide web is te belangrijk geworden om verloren te laten gaan. In korte tijd is het web tot een onmisbare pijler voor informatieverstrekking en dienstverlening geworden. Steeds vaker wordt informatie enkel nog via internet gepubliceerd. Dankzij de laagdrempeligheid van het publiceren en vinden van informatie is het tot één van de belangrijkste dragers van onze hedendaagse cultuur geworden. Het belang hiervan neemt alleen maar toe.**

De crawler van de zoekmachine Google herkende in het jaar 2000 1 miljard pagina's; in 2008 was dit gestegen naar 1000 miljard.[1] Begin 2012 werd er op de videosite YouTube iedere seconde maar liefst één uur aan beeldmateriaal geüpload.[2]
Een keerzijde van deze laagdrempeligheid is de vluchtigheid van op het web gepubliceerd materiaal. Pagina's en complete websites veranderen voortdurend of verdwijnen compleet. Volgens schattingen verdwijnt een gemiddelde webpagina na 100 dagen.[3] Ook informatie waarvan je het zo niet direct zou verwachten, zoals de op het web gepubliceerde persberichten van het Witte Huis, blijkt voortdurend te veranderen.[4] Vaak is informatie op het internet uniek; als de sitebeheerder een pagina om wat voor reden dan ook verwijdert of aanpast, is informatie vaak voorgoed weg. Hiermee gaat in potentie belangrijk cultureel erfgoed verloren: onderzoeksmateriaal voor huidige en toekomstige onderzoekers, bronnen voor verantwoording en materiaal dat economische waarde vertegenwoordigt. "Internet dreigt een gat in de geschiedschrijving te veroorzaken", aldus Amerikaanse historici op een congres in 1998.[5] Bijna 15 jaar later is er nog steeds alle reden voor die zorg, zo niet veel meer.

## Het archiveren van websites

Het is mogelijk om websites of webpagina's te bewaren zodat het voortbestaan niet meer afhankelijk is van een eventueel archief op de betreffende site zelf. De meest gangbare methode voor het aanleggen van zo'n webarchief wordt *webharvesting* of ook wel *webarchivering*[6] genoemd. Door middel van webharvesting kunnen websites als het ware 'bevroren' worden tot een soort van 'digitale prints', dit met behoud van een groot deel van de oorspronkelijke eigenschappen.
Het is niet eenvoudig webarchivering goed uit te voeren. Alleen al het Nederlandse web is enorm groot en verandert voortdurend. Alles bewaren is dus niet mogelijk, selectie is hoe dan ook vereist. Dat creëert een lastig probleem. Een webarchief wordt in essentie meer voor de toekomstige dan voor de huidige gebruikers aangelegd. Hoe kunnen we nu weten in welke informatie toekomstige gebruikers straks het meest geïnteresseerd zijn? Bovendien is de technologie weerbarstig, zowel de technologie van het web zelf als die van de benodigde



*Screenshot Web Curator Tool.*

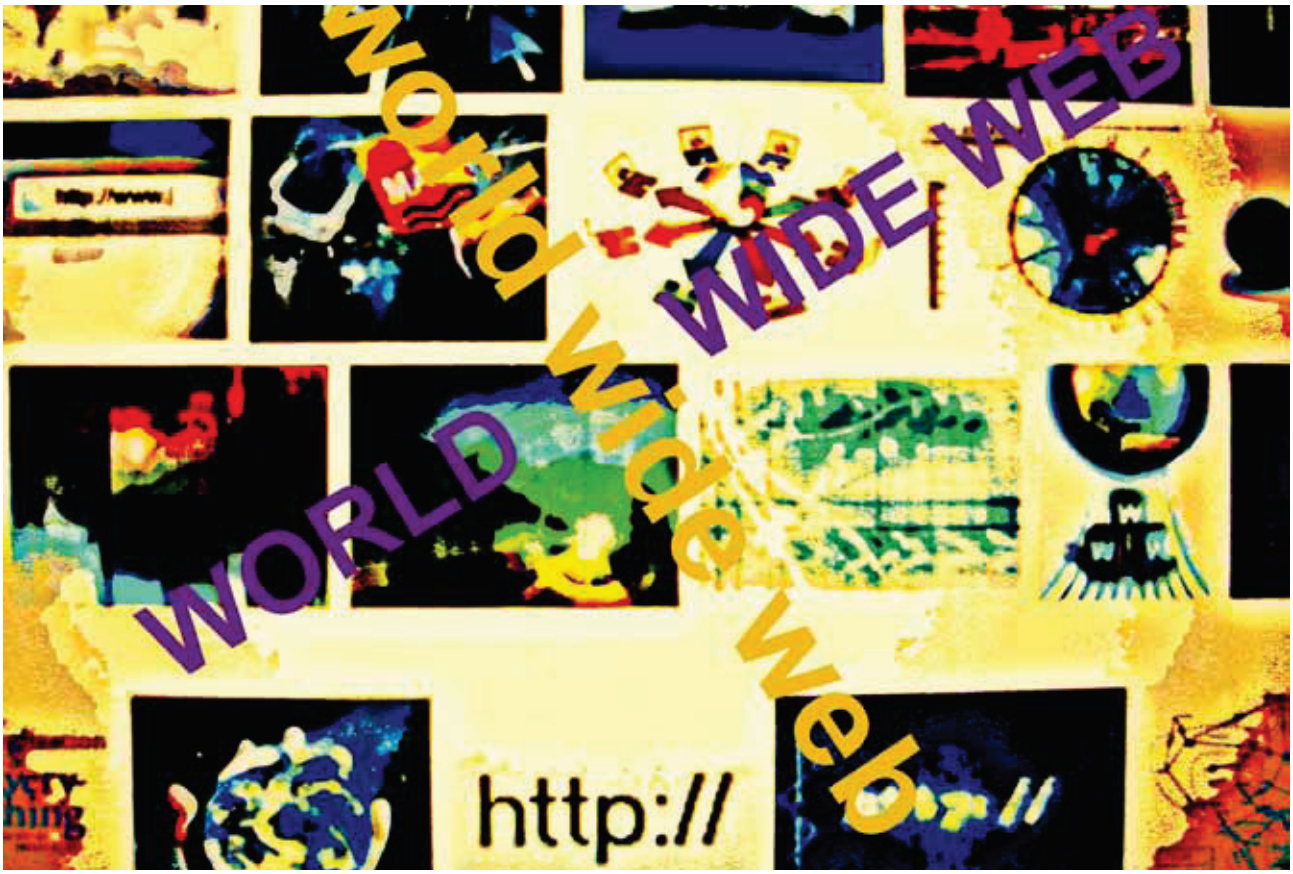
hulpmiddelen voor harvesting, duurzame opslag en toegang.
Webharvesting is dus een tamelijk ingewikkelde activiteit, zowel op het gebied van selectie als van de technische uitvoering. Het is aannemelijk dat juist de kleinere en meer specialistische organisaties binnen hun expertisegebied tot het best onderbouwde en meest toekomstvaste selectiebeleid kunnen komen. Juist voor deze kleinere organisaties zal het weer lastig zijn om de technische drempels te nemen.

## De huidige situatie

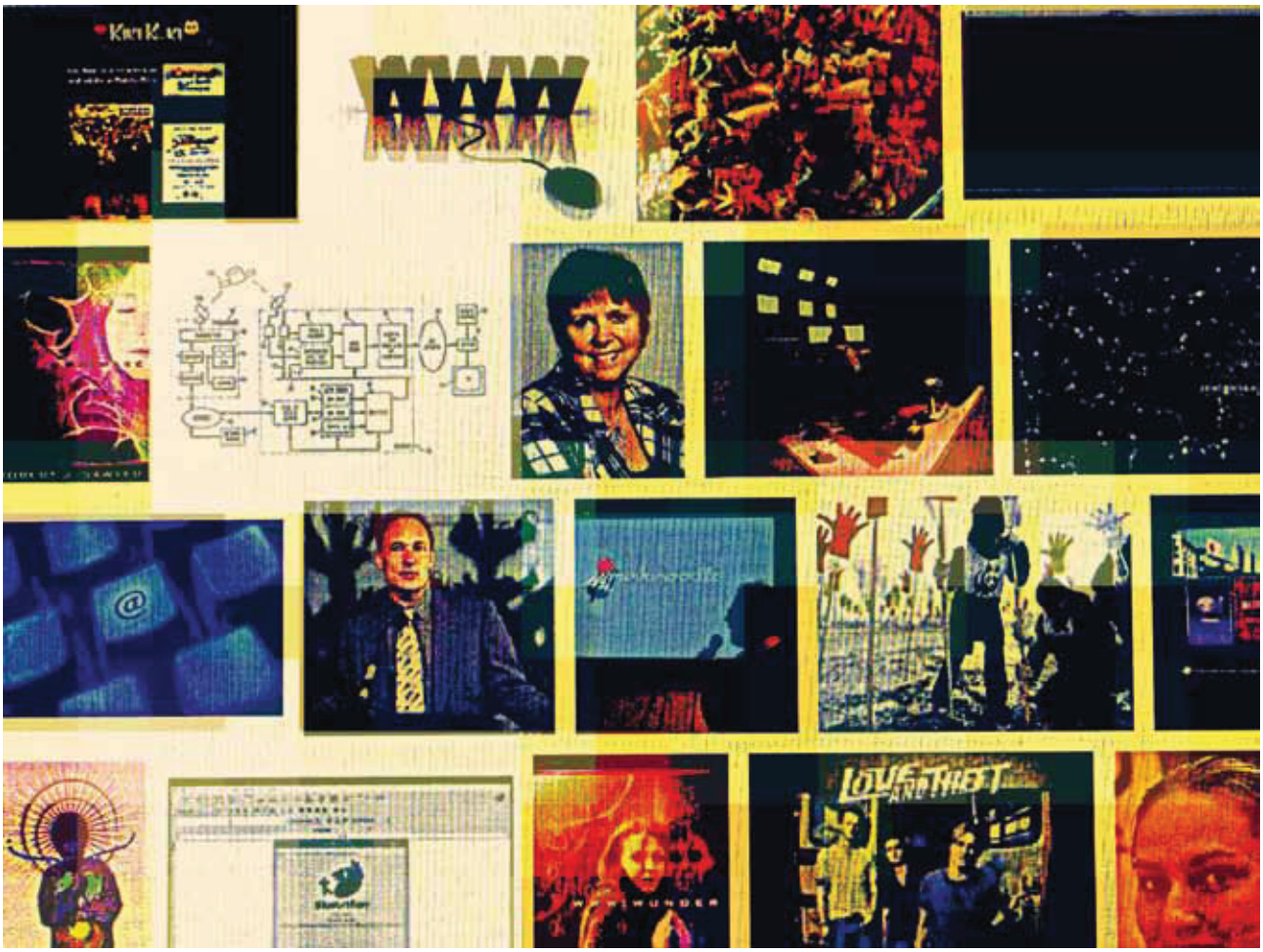
Het Internet Archive (IA) in de Verenigde Staten is één van de pioniers van webarchivering. Het webarchief van het IA, dat ook online beschikbaar is, bevat 'harvests' van websites vanaf 1996 tot heden, ook van het Nederlandse web. Het IA volgt hierbij een aanpak die 'bulkarchivering' genoemd wordt. >>

>> Kort gesteld komt het erop neer dat wordt geprobeerd een paar keer per jaar een zo breed mogelijke selectie van websites te harvesten, ongeveer op de wijze waarop Google het complete internet tracht te indexeren. Door de omvang van het web is het echter onmogelijk om alles te archiveren. Het IA kiest er daarom voor om zoveel mogelijk verschillende sites op te slaan, maar geeft daarbij minder prioriteit aan het zo volledig mogelijk harvesten van die sites. Ook mist het IA wel eens een site, die zo dus helemaal niet gearchiveerd wordt.
Internationaal zijn het doorgaans de nationale bibliotheken die vanuit zogeheten depotwetgeving het nationale domein archiveren. Ze hebben een wettelijk vastgestelde taak om alle publicaties uit het nationale domein te bewaren. In Nederland ontbreekt deze wetgeving en is er auteursrechtelijke toestemming nodig voor het archiveren van websites. Dit maakt bulkarchivering in Nederland praktisch niet uitvoerbaar. Mede door deze lastige uitgangssituatie is de Koninklijke Bibliotheek (KB) pas in 2007 begonnen met de archivering van een handmatige selectie van Nederlandse websites. Deze selectieve aanpak legt de focus op het zo volledig mogelijk archiveren van websites, in tegenstelling tot de genoemde brede aanpak van het IA. Bij iedere site wordt om toestemming gevraagd deze te mogen archiveren en beschikbaar te stellen. Momenteel omvat deze selectie ongeveer 4.000 websites over de Nederlandse maatschappij, cultuur en geschiedenis (met een totale omvang van ongeveer 7 TB). Er wordt gewerkt aan een verbreding van deze selectie.
Een belangrijk aandachtspunt van de KB is onderzoek naar duurzame opslag en toegankelijkheid. Hiertoe wordt wereldwijd met diverse organisaties samengewerkt, zoals het International Internet Preservation Consortium (IIPC). Veel van de door de KB gebruikte hulpmiddelen, zoals de Web Curator Tool (WCT), zijn producten van internationale samenwerking. Nationaal, in het kader van het NWO-onderzoeksprogramma CATCH, wordt nu in het project Web Archive Retrieval Tools (WebART)[7] een 4-jarig onderzoek gedaan naar informatietools en -methoden om het webarchief optimaal toegankelijk te maken voor onderzoekers.
De Nederlandse pionier op het gebied van webarchiveren is het Documentatiecentrum Nederlandse Politieke Partijen (DNPP) van de Rijksuniversiteit Groningen, dat in 2001 startte met het project Archipol (Archiveren websites politieke partijen).[8] Het Documentatiecentrum signaleerde dat de websites van de partijen veel materiaal bevatten dat vroeger in gedrukte vorm verscheen, maar nu niet meer (zoals bijvoorbeeld congresstukken). Hierdoor zouden de 'traditionele' collecties van gedrukt partijmateriaal van het DNPP steeds grotere hiaten gaan vertonen. Ook werden de sites van de partijen steeds belangrijker voor communicatie met leden en kiezers. Voor toekomstig wetenschappelijk onderzoek was (en is) het daarom van groot belang de sites als bronnenmateriaal te bewaren. De speciaal ontwikkelde software die het DNPP voor webarchivering gebruikte, is na tien jaar verouderd. In het archief zitten bijna 1.000 verschillende websites (met een totale omvang van 642 GB), waarvan bijna de helft nog steeds wordt gearchiveerd. Met een in het kader van het project 'Digitaliseren met Beleid' >>

> "Internet dreigt een gat in de geschiedschrijving te veroorzaken."




*Het web is tot één van de belangrijkste dragers van onze hedendaagse cultuur geworden. De keerzijde hiervan is echter dat door de vluchtigheid van het medium potentieel belangrijk cultureel erfgoed verloren gaat.*




*Door middel van webarchivering kunnen websites als het ware 'bevroren' worden tot een soort 'digitale prints', dit met behoud van een groot deel van de oorspronkelijke eigenschappen.*

>> verleende subsidie is het DNPP overgestapt op WCT, dat zoals vermeld ook door de KB wordt gebruikt[9]. Aan dit besluit is intensief overleg met de KB voorafgegaan. Van grote invloed bij de keuze was het inzicht dat het voor een kleine instelling als het DNPP vele voordelen biedt om een grotere speler als de KB te kunnen volgen bij de technologische vernieuwingen wat betreft het downloaden, bewaren, beheren en beschikbaar stellen van gearchiveerde websites.
Naast het DNPP en de KB hebben ook andere instellingen het initiatief genomen om selecties van websites te archiveren. Zo experimenteert het Nederlands Instituut voor Beeld en Geluid met het archiveren van websites van omroepen en webgebaseerde media. Ook zijn er enkele lokale initiatieven, zoals een cultuurhistorisch archief van Rotterdamse websites dat wordt bijgehouden door het Gemeentearchief Rotterdam.
Al deze initiatieven zijn vooral gefocust op collectievorming (het verzamelen van verschillende websites vanuit het perspectief van erfgoed en onderzoek). Een andere vorm van webarchivering, die vooral voor overheidsorganisaties van belang is, is het archiveren van de eigen website vanuit het kader gesteld door de Archiefwet. Voor overheidsorganisaties, en andere zorgdragers inzake de Archiefwet, is het van belang dat ook websites tot 'archief' gerekend kunnen worden. Of websites voor permanente bewaring in aanmerking komen, wordt bepaald door (interpretatie van de) selectielijsten. In de praktijk gebeurt het nu nog slechts mondjesmaat dat zorgdragers ook hun website archiveren of laten archiveren. Dikwijls wordt er gebruik gemaakt van diensten van commerciële partijen, zoals ArchiefWeb, PageFreezer en Capsis.

## Wat zou wenselijk zijn

Verspreid over het land zijn er diverse initiatieven ondernomen om Nederlandse websites te archiveren. Op die manier is een (weliswaar bescheiden deel) van het digitale culturele erfgoed veilig gesteld. Daar staat tegenover dat men vaak niet van elkaars bestaan weet, en potentiële gebruikers nog minder. Een eenduidige structuur om het webarchiveren op gecoördineerde wijze aan te pakken, waarbij men kan leren van elkaars mislukkingen en successen, bestaat nog niet. Wel is duidelijk dat het Nederlandse deel van het internet technisch gezien nooit volledig vanuit één punt kan worden gearchiveerd, wat de vraag oproept wat dan wel het meest ideale organisatorische model van webarchiveren is. In 1996 verscheen in de Verenigde Staten het rapport *Preserving Digital Information* van de Commission on Preservation and Access en The Research Libraries Group. Hierin werd bepleit geen gecentraliseerde structuur op te zetten, maar een decentraal netwerk van instellingen die de archivering van digitale objecten zoals websites, hun bewaring en beschikbaarstelling als taak hebben. Het rapport benadrukt het belang van een decentrale opzet, want "a distributed structure, built on a foundation of electronic networks, places archival responsibility with those who presumably care most about and have the greatest understanding of the value of particular digital information objects".[10]
Een dergelijke structuur zouden wij hier ook willen bepleiten: een netwerk van instellingen in Nederland die websites >>

>> archiveren die gezien hun doelstellingen op hun werkterrein liggen. Deze instellingen beschikken immers over de inhoudelijke kennis welke sites moeten worden gearchiveerd, en hoe vaak.
Een Nederlandse variant van dit 'Amerikaanse' model voor webarchivering zou kunnen zijn dat een partij als de Nationale Coalitie Digitale Duurzaamheid (NCDD) een coördinerende rol op zich neemt. Een taak daarbinnen zou bestaan uit het inzichtelijk maken van wat er van het Nederlandse web gearchiveerd wordt en door wie. Ook zou de coördinerende organisatie een rol moeten spelen in het inzichtelijk maken van de kennisbehoefte van zowel de huidige spelers als ook van de partijen die nu nog niet aan webarchivering doen, omdat het hen aan de nodige kennis of infrastructuur ontbreekt. Een eerste aanzet hiertoe heeft de NCDD vorig jaar al gegeven door een rondetafelgesprek te organiseren waar ook de vraag werd voorgelegd hoe webarchivering in Nederland beter georganiseerd kan worden. In lijn met ons betoog werd geconcludeerd dat kleinere instellingen behoefte hebben aan kennis, in de vorm van handvatten en praktische voorbeelden.[11]
Zeker gezien de aankomende fusie tussen KB en het Nationaal Archief (NA) zouden wij verder willen pleiten voor een prominente inhoudelijke rol binnen dit netwerk voor de 'KBNA'. Daarbij zijn dan in praktijk verschillende scenario's denkbaar.
In de eerste plaats zou KBNA als een soort kenniscentrum kunnen fungeren op het gebied van webarchivering, waar relevante kennis wordt verzameld en gedeeld. Denk bijvoorbeeld aan uitgewerkte casestudy's, maar ook aan richtlijnen voor de bouw van archiveerbare websites. Als lid van het International Internet Preservation Consortium (IIPC) is de KB van de laatste technologische ontwikkelingen op de hoogte, en die informatie kan zij aan de 'decentrale' instellingen doen toekomen. Daarbij zou KBNA ook als laagdrempelige technologische vraagbaak kunnen dienen.
Dit scenario is niet zo ingrijpend en is dunkt ons mogelijk zonder al te veel extra financiële kosten. Een verdergaande variant is het model waarin KBNA de technologische support en structuur (al dan niet inclusief een voorziening voor de opslag van de gearchiveerde websites) aanbiedt aan de decentrale instelling, die zelf wel verantwoordelijk is voor de selectie van de voor archivering in aanmerking komende websites. Aan deze oplossing zitten veel meer haken en ogen: van wie 'zijn' de gearchiveerde websites?; vindt er een kostenverrekening plaats en zo ja, op welke wijze?; wat moet er gebeuren als de site maar in beperkte mate aansluit bij het selectieprofiel van KBNA?

> Het is een idee fixe om te denken dat alle websites bewaard zouden moeten worden.

## Afsluitend

Het is zonder meer positief dat er in Nederland al op een aantal plaatsen begonnen is met het archiveren van websites, soms al zo'n tien jaar lang. Tegelijk moet worden vastgesteld dat deze initiatieven een druppel zijn op een gloeiende plaat. De groei van het internet gaat zo snel en het internet zelf wordt zo divers (bijvoorbeeld het archiveren van social network sites als Facebook, Twitter en dergelijke brengt allerlei nieuwe technische en praktische uitdagingen met zich mee), dat er sprake is van een steeds grotere achterstand. Het is een idee fixe om te denken dat alle websites bewaard zouden moeten worden – net als in het werkelijke leven is niet alles de moeite waard om te archiveren. Een scherpe selectie is geboden, en die zou kunnen worden uitgevoerd door erfgoedinstellingen in de brede zin des woords, die elk op hun eigen specifieke terrein de kennis in huis hebben van wat 'bewaar-waardig' is. Idealiter zouden zij daarbij door een grote speler als KBNA, met haar internationale netwerken, technologisch ondersteund moeten worden, omdat zij over die kennis juist zelf vaak niet beschikken. Een decentraal netwerk rondom een faciliterende nationale instantie – een dergelijke structuur lijkt ons wenselijk om het webarchiveren in Nederland een flinke impuls te geven. ■

## Noten

**1** ■ http://googleblog.blogspot.nl/2008/07/we-knew-web-was-big.html, geraadpleegd 16 juli 2012.
**2** ■ http://youtube-global.blogspot.nl/2012/01/holy-nyans-60-hours-per-minute-and-4.html, geraadpleegd 16 juli 2012.
**3** ■ Zie http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage, geraadpleegd 16 juli 2012.
**4** ■ http://www.clinecenter.illinois.edu/airbrushing_history/.
**5** ■ *De Volkskrant*, 30 januari 1999.
**6** ■ De term 'webharvesting' heeft specifiek betrekking op het vergaren van materiaal op het web, bijvoorbeeld om het duurzaam te kunnen bewaren. 'Webarchivering' is een wat meer omvattende term die dikwijls gebruikt wordt voor het geheel aan processtappen benodigd voor het duurzaam bewaren en toegankelijk maken van webpagina's en -sites.
**7** ■ http://www.nwo.nl/catch/webart, geraadpleegd 16 juli 2012.
**8** ■ Dit project werd gesubsidieerd door de Rijksuniversiteit Groningen en de stuurgroep Innovatie Wetenschappelijke Infor¬matievoorziening. Zie onder meer A.K. Keijzer, F.J. den Hollander en G. Voerman, 'Het Archipol-project. Het archiveren van websites van Nederlandse politieke partijen', in: *Archievenblad* 106 (2002) nr. 1 (feb.), 32-33; Frank den Hollander en Gerrit Voerman, *Het web gevangen. Het archiveren van websites van Nederlandse politieke partijen* (Groningen 2002).
**9** ■ Voor geïnteresseerde instellingen is deze open source-software beschikbaar.
**10** ■ *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group* (z.pl., 1996) 21; zie www.clir.org/pubs/reports/pub63watersgarrett.pdf.
**11** ■ http://www.ncdd.nl/blog/?p=129.

Gerrit Voerman ■ Documentatiecentrum Nederlandse Politieke Partijen, Rijksuniversiteit Groningen.
René Voorburg ■ werkzaam bij de Koninklijke Bibliotheek, bijdrage op persoonlijke titel.
Hugo Huurdeman ■ promovendus aan de Faculteit der Geesteswetenschappen, Universiteit van Amsterdam.